(2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine*, *170*(1), 51-58. https://doi.org/10.7326/M18-1376

PROBAST: A tool to assess the risk of bias and applicability of prediction model studies

Robert F. Wolff, MD[1,#], Prof Karel G. M. Moons, PhD[2,3,#], Prof Richard D. Riley, PhD[4], Penny F. Whiting, PhD[5,6], Marie Westwood, PhD[1], Prof Gary S. Collins, PhD[7], Johannes B. Reitsma, MD, PhD[2,3], Prof Jos Kleijnen, MD, PhD[1,8], Sue Mallett, DPhil[9] on behalf of the PROBAST group[*]

[1] Kleijnen Systematic Reviews Ltd, York, United Kingdom

[2] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

[3] Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands

[4] Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Keele, United Kingdom

[5] Bristol Medical School, University of Bristol, Bristol, United Kingdom

[6] NIIHR CLAHRC West, University Hospitals Bristol NHS Foundation Trust, Bristol, United Kingdom

[7] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Diseases, University of Oxford, Oxford, United Kingdom

[8] School for Public Health and Primary Care (CAPHRI) Maastricht University, Maastricht, The Netherlands

[9] Institute of Applied Health Research, NIHR Birmingham Biomedical Research Centre, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom

[#] Both authors contributed equally

[*] All members of the PROBAST group are listed in the appendix

Corresponding author:
Dr Robert Wolff
Kleijnen Systematic Reviews Ltd
Unit 6
Escrick Business Park
Riccall Road
Escrick
York YO19 6FD
United Kingdom
Tel. +44 (0)1904 727987
Fax. +44 (0)1904 720429
Email. robert@systematic-reviews.com

Short title: PROBAST
Word count: 3,529 words (Introduction; Methods; Results; Discussion)

Keywords: Bias (Epidemiology); Diagnosis, Evidence-Based Medicine; Multivariable Analysis; Prediction; Prognosis; Reproducibility of Results

Abstract

(199 words)

Clinical prediction models combine multiple predictors to estimate the risk of whether a particular condition is present (diagnostic) or whether a certain event will occur in the future (prognostic).

PROBAST, a tool for assessing the risk of bias (ROB) and applicability of diagnostic and prognostic prediction model studies, considered existing ROB tools as well as reporting guidelines and was developed by a steering group, informed by a Delphi procedure involving 38 experts and refinement through piloting.

PROBAST is grouped into four domains: participants, predictors, outcomes, and analysis. These domains contain a total of twenty signalling questions to facilitate structured judgement of ROB. We define ROB to occur when shortcomings in study design, conduct or analysis lead to systematically distorted estimates of model predictive performance. PROBAST enables a focussed and transparent approach to assessing the ROB and applicability of studies developing, validating or updating prediction models for individualised predictions.

Although PROBAST was designed for use in systematic reviews, it can be used more generally in critical appraisal of prediction model studies. Potential users include organisations supporting decision making, researchers and clinicians with an interest in evidence-based medicine or involved in guideline development as well as journal editors and manuscript reviewers.

Introduction

(546 words)

Prediction relates to estimating the probability of something currently unknown. In the context of medical research, prediction typically relates to either diagnosis (probability of a certain condition being present but not yet detected) or prognosis (probability of developing a future outcome).(1-3) Prognosis does not only apply to sick individuals or those with an established diagnosis, but also to, for example, prognosis of pregnant women at risk of developing diabetes.(4) Prediction research includes predictor finding studies, prediction model studies (development, validation and extending or updating), and prediction model impact studies.(1)

Predictor finding studies (also known as risk factor or prognostic factor studies) aim to identify which predictors (e.g. age, disease stage, biomarkers) independently contribute to the prediction of a diagnostic or prognostic outcome.(1, 5)

Prediction model studies typically aim to develop, validate or update (e.g. extend) a multivariable prediction model. In a prediction model, multiple predictors are used in combination for estimating probabilities to inform and often guide individual care.(2, 6, 7) These models can either predict an individual's probability of currently having a particular outcome or disease (diagnostic prediction model) or experiencing a particular outcome in the future (prognostic prediction model). Prediction models, both diagnostic and prognostic, are widely used for a variety of medical domains and settings,(8-10) evidenced by the large number of models developed, especially in cancer,(11, 12) neurology,(13, 14) and cardiovascular disease domains.(15) Prediction models are sometimes described as risk prediction models, predictive models, prediction indices or rules, or risk scores.(2, 7) An example is QRISK2 for predicting cardiovascular risk.(16)

Prediction model impact studies evaluate the effect of using a model to guide patient care compared to not using such a model, and focus on the effect of its use on clinical decision making, patient outcomes, or costs of care, using a comparative design such as a randomised trial.(1)

Systematic reviews have a key role in evidence-based medicine and in the development of clinical guidelines.(17-19) They are considered to provide the most reliable form of evidence for the effects of an intervention or diagnostic test.(20, 21) Systematic reviews of prediction models are a relatively new and evolving area but are increasingly undertaken to systematically identify, appraise and summarise evidence on the performance of prediction models.(1, 6, 22)

Quality assessment of included studies is a crucial step in any systematic review.(20, 21) The QUIPS tool has been developed to assess the risk of bias (ROB) in predictor finding (prognostic factor) studies.(23) The methodological quality of studies investigating the impact of a prediction model using a comparative randomised design can be assessed using the revised Cochrane ROB tool (ROB 2.0)(24) or ROBINS-I for non-randomised comparative designs.(25) With the increased number of prediction model studies as well as systematic reviews of prediction model studies, a tool facilitating quality assessment for individual prediction model studies is urgently needed.

We present PROBAST (Prediction model Risk Of Bias ASsessment Tool), a tool to assess the ROB and concerns regarding the applicability of diagnostic and prognostic prediction model studies. PROBAST can be used to assess both model development and model validation studies, including those updating a prediction model (Box 1). We explicitly refer to the accompanying Explanation and Elaboration paper

for detailed explanations on how to use the PROBAST tool and how to make ROB and applicability judgements.[REF M18-1377]

Methods – Development of PROBAST

(813 words)

Development of PROBAST was based on a four-stage approach for developing health research reporting guidelines: define the scope, review the evidence base, web-based Delphi procedure, and refine the tool through piloting.(27) Guidelines explicitly aimed at the development of quality assessment tools were not available at the time.(28)

Development stage 1: Scope and Definitions

A steering group of nine experts in the area of prediction model studies and quality assessment tool development agreed on key features of the desired scope of PROBAST. The scope was further refined during the web-based Delphi procedure with a panel of 38 experts with different backgrounds.

PROBAST was primarily designed to assess primary studies included in a systematic review. The group agreed that PROBAST would assess both, the *risk of bias* and *concerns regarding applicability,* of a study evaluating a multivariable diagnostic or prognostic prediction model to be used for individualised predictions. A domain-based structure was adopted similar to that used in other risk of bias tools such as ROB 2.0,(24) ROBINS-I,(25) QUADAS-2,(29) and ROBIS.(30)

It was agreed that PROBAST should cover primary studies that developed, validated or updated one or more multivariable prediction models for the purpose of making individualised predictions of a diagnostic or prognostic outcome (Box 1). Studies using multivariable modelling techniques to identify predictors (e.g. risk or prognostic factors) associated with an outcome but not attempting to develop, validate or update a model for making individualised predictions are not covered by PROBAST.(5) Therefore, PROBAST is not intended for predictor finding studies and prediction model impact studies.

Diagnostic and prognostic model studies often use different terms for the predictors and outcomes (Box 2). A multivariable prediction model is defined as any combination or equation of two or more predictors for estimating the probability or risk for an individual.(6, 7, 31-33)

Development stage 2: Review of Evidence

Three different approaches were used to provide an evidence base to inform the development of PROBAST: (1) identification of relevant methodological reviews in the area of prediction model research (November 2012 to January 2013), (2) asking members of the steering group to identify relevant methodological studies (January 2013 to March 2013), and (3) use of the Delphi procedure to ask members of the wider group to identify additional evidence (February 2012 to July 2014).

Identified literature was used to guide the scope and produce an initial list of signalling questions for consideration for inclusion in PROBAST.(1, 2, 5-7, 26, 32-39) Signalling questions were grouped into common themes in order to identify possible domains. Additional literature provided as part of the web-based surveys was used to inform the development of the E&E paper.

Development stage 3: Web-based Delphi procedure

A modified Delphi process was used to gain feedback and agreement on the scope, structure and content of PROBAST. Web-based surveys were developed to gather structured feedback for each round. The Delphi group included 38 members comprising methodological experts in the areas of prediction model research and quality assessment tool development, experienced systematic reviewers, commissioners, and representatives of reimbursements agencies. Different potential

stakeholders were included to ensure that the views of end-users, methodological experts and decision makers were represented.

The Delphi process consisted of seven rounds. Round 1 asked about the scope of the tool and it was agreed to focus on prediction model studies only and to follow a domain-based structure. Round 2 aimed at identifying and finding a consensus regarding the relevant domains to be included. The signalling questions for domains were refined in rounds 3 to 5. Respondents were asked to rate each proposed signalling question for inclusion using a 1 to 5 Likert scale. They were also given the opportunity to provide suggested rephrasing, provide any supporting evidence (e.g. references to relevant studies) and suggest any missing signalling questions. Round 6 refined the domains and introduced further optional guidance for the use of PROBAST. In the last round, participants were sent the agreed draft version of PROBAST and given the opportunity to provide any final feedback.

Development stage 4: Piloting and refining of the tool

Six workshops on PROBAST were held at consecutive annual Cochrane Colloquia (Quebec 2013, Hyderabad 2014, Vienna 2015, Seoul 2016, Cape Town 2017, Edinburgh 2018) and numerous consecutive workshops with MSc and PhD students (e.g. MSc Epidemiology program of Utrecht University, The Netherlands, and Evidence Based Health Care program of Oxford University, UK). In these, we piloted the then current version of the PROBAST tool to gather feedback on the practical issues associated with using the tool so we could further refine and subsequently validate the tool. Finally, over fifty review groups have already piloted PROBAST versions, included the final version, in their reviews. Topics included cancer, cardiology, endocrinology, pulmonology and orthopaedics.

All feedback received from these initiatives was used to further inform the content and structure of the PROBAST tool, wording of the signalling questions, and content of the guidance documents.[REF M18-1377]

Results – The PROBAST tool

(1,640 words)

What does PROBAST assess?

PROBAST assesses both the *risk of bias* and *concerns regarding applicability* of primary studies that developed or validated one or more multivariable prediction models for diagnosis or prognosis (Boxes 1 and 2).

Development of a prediction model can include adding new predictors to an existing prediction model. Similarly, validation of an existing model can be accompanied by updating and also extending of the model, i.e. the development of a new model. PROBAST is applicable to both situations (Box 1).

Target users

Although PROBAST was designed for use in systematic reviews, it can be used more generally in critical appraisal of prediction model studies. Potential users of PROBAST include organisations supporting decision making (e.g. National Institute for Health and Care Excellence, NICE; Institute for Quality and Efficiency in Health Care, IQWiG), researchers and clinicians with an interest in evidence-based medicine or involved in guideline development as well as journal editors, manuscript reviewers and readers wanting to critically appraise prediction model studies.

Definition of risk of bias and applicability

Bias is usually defined as presence of systematic error within a study leading to distorted or flawed study results, hampering the internal validity of that study. In prediction model development and validation, there are known features which make a study at ROB, although there is limited *empirical* evidence to demonstrate the most important sources of bias. We define risk of bias to occur when shortcomings in the study design, conduct or analysis lead to systematically distorted estimates of model predictive performance. Model predictive performance is typically evaluated using measures of calibration and discrimination, and sometimes (notably in diagnostic model studies) classification.(7) To understand bias in study estimates of model predictive performance, it helps to think about how a hypothetical methodologically robust prediction model study would have been designed, conducted and analysed. Many sources of bias identified in other medical research areas are also relevant to prediction model studies, such as blinding of assessors of study outcomes to other features of the study, and the use of consistent definitions and measurements for predictors and outcomes within the study.

Concerns regarding the applicability of primary studies to the review question can arise when the study population, predictors or outcomes of a primary study differ from those specified in the review question. Applicability concerns may arise when participants in the prediction model study are from a different medical setting than the population defined in the review question. For example, participants in a primary prediction model study may be enrolled from a hospital setting but the review question specifically relates to participants in primary care. The reported prediction model discrimination and calibration may not be applicable, as patients in hospital settings typically have more severe disease than patients in primary care.(40, 41)

For systematic reviews where eligibility criteria, predictors and outcomes of the primary studies, directly match the review question, there will be no concerns for applicability of a primary study for the review. However, typically systematic reviews have inclusion criteria that are broader than the

focus of the review question. The broader inclusion criteria allow for variation in the searching of the primary studies and thus require careful assessment of applicability of each primary study to the actual review question.(7)[REF M18-1377]

Types of prediction model study

A primary study identified as relevant for the review may include the development, validation or update of one or more prediction models. For each study, a PROBAST assessment should be completed for each distinct model that is developed, validated, or updated for making individualised predictions, relevant to the systematic review question.

PROBAST includes four steps (Table 1). We stress the importance of the accompanying paper which provides detailed explanations and guidance for completing each step.[REF M18-1377]

Step 1: Specify your systematic review question

Assessors are first asked to report their systematic review question in terms of intended use of the model, targeted participants, predictors used in the modelling, and predicted outcome. Specific guidance (i.e. the CHARMS checklist) exists to help reviewers define a clear and focused review question.(22, 26)

Step 2: Classify the type of prediction model evaluation

Different signalling questions apply for different types of prediction model evaluation. For each model assessment, reviewers classify a model as “development only”, “development and validation in the same publication” or “validation only”. When a publication focuses on creating a model by adding one or more new predictors to established predictors (or an established model), “development only” should be used. When a publication focuses on validation of an existing model in other data though followed by updating (adjusting or extending) of the model such that in fact a new model is being developed, then “development and validation in the same publication” should be used. Note again that sometimes a single publication may address more than one model of interest.

Step 3: Assess risk of bias and applicability

Step 3 aims to identify areas where bias may be introduced into the prediction model study or where there may be concerns for applicability. It involves the assessment of four domains to cover key aspects of prediction model studies: (1) participants, (2) predictors, (3) outcome, and (4) analysis. The risk of bias component of each domain comprises four sections: information used to support the judgment, 20 signalling questions (2 to 9 per domain), judgment of ROB, and rationale regarding the judgment (Table 2)

The support for judgement box provides space to record the information used to answer the signalling questions. Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). Risk of bias is judged as “low”, “high”, or “unclear”. All signalling questions are phrased so that “yes” indicates absence of bias. Any signalling question rated as “no” or “probably no” flags the potential for bias; assessors will need to use their own judgment to determine whether the domain should be rated as “high”, “low” or “unclear” risk of bias. A “no” rating does not automatically result in a “high” risk of bias rating. The “no information” category should be used only when insufficient information is reported to permit a judgment. By recording the rationale for the risk of bias rating, the rating will be transparent and, where necessary, facilitate discussion among review authors completing assessments independently.

The first three domains are also rated for concerns regarding applicability (low / high / unclear) to the review question defined above. Concerns regarding applicability are rated in a similar way to ROB but there are no signalling questions.

All domains should be completed separately for each evaluation of a distinct model in each study. The team completing a PROBAST assessment is likely to need both subject content and methodological expertise to complete an assessment. For further details on how to score ROB and applicability concerns we refer to the accompanying paper and www.probast.org.[REF M18-1377]

- **Domain 1 (Participants)** covers potential sources of bias and applicability concerns related to how participants were selected for enrolment into the study and the data sources (e.g. study designs) used. Two signalling questions support the assessment of risk of bias.
- **Domain 2 (Predictors)** covers potential sources of bias and applicability concerns related to the definition and measurement of the predictors evaluated for inclusion in the prediction model. Three signalling questions support the assessment of risk of bias.
- **Domain 3 (Outcome)** covers potential sources of bias and applicability concerns related to the definition and measurement of the outcome that is predicted by the model. Six signalling questions support the assessment of risk of bias.
- **Domain 4 (Analysis)** covers potential sources of bias regarding the statistical analysis methods. It assesses aspects related to the choice of analysis method and whether key statistical considerations (e.g. in regards to missing data) were correctly addressed. Nine signalling questions support the assessment of risk of bias.

Table 2 presents an overview of step 3. Detailed examples how to rate signalling questions and judge domains can be found in the E&E publication and on www.probast.org.[REF M18-1377]

Step 4: Overall judgement

Based on the risk of bias classifications for each domain in step 3, an *overall* judgement about ROB of the prediction model should be made. An overall rating of either low, high, or unclear ROB should be used. We recommend rating the prediction model to be of a low ROB if no relevant shortcomings were identified in the risk of bias assessment, i.e. all domains were rated as “low risk of bias”. If at least one domain was judged to be of high ROB, an overall judgement of high ROB should be used. Similarly, unclear ROB should be assigned once an unclear ROB was noted in at least one domain and it was low risk for all other domains.

However, if a prediction model was developed without any external validation on different participants and even all four domains were rated as low ROB, downgrading to high ROB should still be considered unless the model development was based on a very large data set or included some form of internal validation. For details we refer to the E&E paper.[REF M18-1377]

Based on the applicability classifications for each domain in step 3, an overall judgement about the concerns regarding applicability of the prediction model is needed. A “low concern” decision should only be reached if all domains showed low concerns regarding applicability. Similarly, if one or more domains were judged to have high concerns regarding applicability, the overall judgement should be “high concern”. “Unclear concerns regarding applicability” should only be reached if one or more domains are judged as “unclear” regarding applicability and all other domains were rated to have “low concerns”.

Detailed explanation and examples on how to judge the overall ROB and concerns regarding applicability can be found in the accompanying publication and on www.probast.org.[REF M18-1377] Table 3 suggests a way to present the results of the PROBAST assessments.

Discussion

[331 words]

Assessment of the quality of included studies is an essential component of all systematic reviews and evidence syntheses. Systematic reviews of prediction model studies are a rapidly evolving area.(22) With the increased number of prediction model studies as well as systematic reviews of prediction model studies, a tool facilitating quality assessment for individual prediction model studies is urgently needed. PROBAST is the first rigorously developed tool designed specifically to assess the quality of prediction model studies for development, validation or updating models for both diagnostic and prognostic models, regardless of the medical domain, type of outcome, predictors or statistical technique used.

We adopted a domain based structure similar to that used in other recently developed tools such as the revised Cochrane risk of bias tool (ROB 2.0),(24) QUADAS-2 for diagnostic accuracy studies,(29) ROBINS-I for non-randomised studies,(25) and ROBIS for systematic reviews.(30) All stages of PROBAST development included a wide range of stakeholders with piloting starting with early versions of the tool allowing feedback from direct reviewer experience to be incorporated into the final tool. We feel that these two features have resulted in a tool that is both methodologically sound and user-friendly.

Potential users of PROBAST include systematic review authors, healthcare decision makers, researchers and clinicians with an interest in evidence-based medicine or involved in guideline development as well as journal editors and manuscript reviewers.

Explicit guidance and explanation about how to use PROBAST is provided in the accompanying Explanation & Elaboration (E&E) paper.[REF M18-1377] To understand and use the PROBAST tool, we stress that this E&E paper should always be read in conjunction with the current paper. A multidisciplinary team, combining both subject content and methodological expertise, should be used when assessing prediction model studies.

As with other risk of bias and reporting guidelines in medical research, PROBAST and its guidance will require updating, as methods for prediction model studies develop. We recommend downloading the latest version of PROBAST tool and accompanying guidance, including detailed examples from the website (www.probast.org).

Contact details for all authors

Robert F. Wolff
Kleijnen Systematic Reviews Ltd
Unit 6
Escrick Business Park
Riccall Road
Escrick
York YO19 6FD
United Kingdom
robert@systematic-reviews.com

Karel G. M. Moons
Julius Centre for Health Sciences and Primary Care
UMC Utrecht
Utrecht University
PO Box 85500
3508 GA Utrecht
The Netherlands
K.G.M.Moons@umcutrecht.nl

Richard D. Riley
Centre for Prognosis Research,
Research Institute for Primary Care and Health Sciences
Keele University
Staffordshire ST5 5BG
United Kingdom
r.riley@keele.ac.uk

Penny F. Whiting
NIHR CLAHRC West
University Hospitals Bristol NHS Foundation Trust, Bristol, United Kingdom
School of Social and Community Medicine, University of Bristol, United Kingdom
Whitefriars BS1 2NT
United Kingdom
Penny.Whiting@bristol.ac.uk

Marie Westwood
Kleijnen Systematic Reviews Ltd
Unit 6
Escrick Business Park
Riccall Road
Escrick
York YO19 6FD
United Kingdom
marie@systematic-reviews.com

Gary S. Collins
Centre for Statistics in Medicine, NDORMS, University of Oxford

Botnar Research Centre, Windmill Road
Oxford OX3 7LD
United Kingdom
gary.collins@csm.ox.ac.uk

Johannes B. Reitsma
Julius Centre for Health Sciences and Primary Care
UMC Utrecht
Utrecht University
PO Box 85500
3508 GA Utrecht
The Netherlands
J.B.Reitsma-2@umcutrecht.nl

Jos Kleijnen
Kleijnen Systematic Reviews Ltd, Unit 6, Escrick Business Park, Riccall Road, Escrick, York YO19 6FD, United Kingdom
School for Public Health and Primary Care (CAPHRI) Maastricht University, Maastricht, The Netherlands
jos@systematic-reviews.com

Sue Mallett
Institute of Applied Health Sciences
University of Birmingham
Edgbaston
Birmingham B15 2TT
United Kingdom
s.mallett@bham.ac.uk

Acknowledgements

The authors would like to thank the members of the Delphi panel (see below) for their valuable input. Furthermore, the authors would like to thank all testers, especially Cordula Braun, Johanna A.A.G. Damen, Paul Glasziou, Pauline Heus, Lotty Hooft, and Romin Pajouheshnia, for providing feedback on PROBAST. The authors are grateful to Janine Ross and Steven Duffy for their support in managing the references.

KGM Moons and JB Reitsma gratefully acknowledges financial contribution by the Netherlands Organisation for Scientific Research (ZONMW 918.10.615 and 91208004). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

R Riley is a member of the Evidence Synthesis Working Group funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR) [ProjectNumber 390]. The views expressed are those of the author(s) and not necessarily those of the NIHR, the NHS or the Department of Health.

PF Whiting (time) was supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) West at University Hospitals Bristol NHS Foundation Trust.

GS Collins was supported by the NIHR Biomedical Research Centre, Oxford.

S Mallett is supported by NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham.

This report presents independent research supported by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, or the Department of Health.

Potential Conflicts of Interest

Robert F. Wolff: None to declare

Karel G. M. Moons: None to declare

Richard D. Riley: None to declare

Penny F. Whiting: None to declare

Marie Westwood: None to declare

Gary S. Collins: None to declare

Johannes B. Reitsma: None to declare

Jos Kleijnen: None to declare

Sue Mallett: None to declare

Author Contributions

Conception and design: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

Analysis and interpretation of the data: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

Drafting of the article: R.F. Wolff, K.G.M. Moons, P.F. Whiting, M. Westwood, S. Mallett

Critical revision for important intellectual content: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

Final approval of the article: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

Statistical expertise: K.G.M. Moons, R.D. Riley, G.S. Collins, J.B. Reitsma, S. Mallett

Obtaining of funding: K.G.M. Moons, R.D. Riley, P.F. Whiting, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

Administrative, technical, or logistic support: R.F. Wolff, K.G.M. Moons, J. Kleijnen

Collection and assembly of data: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett

References

1. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Medicine. 2012;9(5):1-12.
2. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Medicine. 2013;10(2):e1001381.
3. Knottnerus JA**.** Diagnostic prediction rules: principles, requirements and pitfalls. Primary Care. 1995;22(2):341-63.
4. Lamain-de Ruiter M, Kwee A, Naaktgeboren CA, de Groot I, Evers IM, Groenendaal F, et al. External validation of prognostic models to predict risk of gestational diabetes mellitus in one Dutch cohort: prospective multicentre cohort study. BMJ. 2016;354:i4338.
5. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Medicine. 2013;10(2):e1001380.
6. Collins GS, Reitsma JB, Altman DG, Moons KG**.** Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
7. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of Internal Medicine. 2015;162(1):W1-73.
8. Collins GS, Mallett S, Omar O, Yu LM**.** Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine. 2011;9:103.
9. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA. 2011;306(15):1688-98.
10. Steurer J, Haller C, Häuselmann H, Brunner F, Bachmann LM**.** Clinical value of prognostic instruments to identify patients with an increased risk for osteoporotic fractures: systematic review. PLoS One. 2011;6(5):e19994.
11. Altman DG**.** Prognostic models: a methodological framework and review of models for breast cancer. Cancer Investigation. 2009;27(3):235-43.
12. Shariat SF, Karakiewicz PI, Suardi N, Kattan MW**.** Comparison of nomograms with other methods for predicting outcomes in prostate cancer: a critical analysis of the literature. Clinical Cancer Research. 2008;14(14):4400-7.
13. Counsell C, Dennis M**.** Systematic review of prognostic models in patients with acute stroke. Cerebrovascular Diseases. 2001;12(3):159-70.
14. Perel P, Prieto-Merino D, Shakur H, Clayton T, Lecky F, Bouamra O, et al. Predicting early death in patients with traumatic bleeding: development and validation of prognostic model. BMJ. 2012;345:e5166.
15. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416.
16. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008;336(7659):1475-82.
17. Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, eds. Clinical practice guidelines we can trust. Washington, DC: National Academies Press; 2011.
18. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014;129(25 Suppl 2):S49-73.
19. Rabar S, Lau R, O'Flynn N, Li L, Barry P**.** Risk assessment of fragility fractures: summary of NICE guidance. BMJ. 2012;345:e3698.
20. Centre for Reviews and Dissemination**.** Systematic Reviews: CRD's guidance for undertaking reviews in health care. York: University of York; 2009.
21. Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions. Chichester: Wiley-Blackwell, The Cochrane Collaboration; 2011.
22. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460.
23. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C**.** Assessing bias in studies of prognostic factors. Annals of Internal Medicine. 2013;158(4):280-6.

24. Higgins JPT, Sterne JAC, Savović J, Page MJ, Hróbjartsson A, Boutron I, et al. A revised tool for assessing risk of bias in randomized trials. In: Chandler J, McKenzie J, Boutron I, Welch V, eds. Cochrane Methods; 2016.
25. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.
26. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Medicine. 2014;11(10):e1001744.
27. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Medicine. 2010;7(2):e1000217.
28. Whiting P, Wolff R, Mallett S, Simera I, Savović J. A proposed framework for developing quality assessment tools. Systematic Reviews. 2017;6(1):204.
29. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine. 2011;155(8):529-36.
30. Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. Journal of Clinical Epidemiology. 2016;69:225-34.
31. Canet J, Gallart L, Gomar C, Paluzie G, Valles J, Castillo J, et al. Prediction of postoperative pulmonary complications in a population-based surgical cohort. Anesthesiology. 2010;113(6):1338-50.
32. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. Journal of Clinical Epidemiology. 2013;66(3):268-77.
33. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375.
34. Harrell FE. Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
35. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:e5595.
36. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Medicine. 2010;8:20.
37. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.
38. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606.
39. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ. 2009;338:b604.
40. Knottnerus JA. Between iatrotropic stimulus and interiatric referral: the domain of primary care research. Journal of Clinical Epidemiology. 2002;55(12):1201-6.
41. Oudega R, Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. Annals of Internal Medicine. 2005;143(2):100-7.

Boxes

Tables

Appendix

The tool is also available on www.probast.org.

PROBAST

(Prediction model study Risk Of Bias Assessment Tool)

PROBAST includes four steps.

Step	Task	When to complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability (per domain)	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment of risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication

If this is your first time using PROBAST, we strongly recommend reading the detailed explanation and elaboration (E&E) paper[REF M18-1377] and to check the examples on www.probast.org.

Step 1: Specify your systematic review question

State your systematic review question to facilitate the assessment of the applicability of the evaluated models to your question. *The following table should be completed once per systematic review.*

Criteria	Specify your systematic review question
Intended use of model:	
Participants *including selection criteria and setting:*	
Predictors (used in modelling) *including (1) types of predictors (e.g. history, clinical examination, biochemical markers, imaging tests), (2) time of measurement, (3) specific measurement issues (e.g. any requirements/ prohibitions for specialised equipment):*	
Outcome *to be predicted:*	

Step 2: Classify the type of prediction model evaluation

Use the following table to classify the evaluation as model development, model validation, or combination. Different signalling questions apply for different types of prediction model evaluation.
When a publication focuses on adding one or more new predictors to established predictors then use "development only". When a publication focuses on validation of an existing model in other data though followed by updating (adjusting or extending) of the model such that in fact a new model is being developed, then use "development and validation in the same publication".
If the evaluation does not fit one of these classifications then PROBAST should not be used.

Classify the evaluation based on its aim			
Type of prediction study	**PROBAST boxes to complete**	**Tick as appropriate**	**Definitions for type of prediction model study**
Development only	Dev		Prediction model development without external validation. These studies may include internal validation methods such as bootstrapping and cross-validation techniques
Development and validation	Dev and Val		Prediction model development combined with external validation in other participants in the same article
Validation only	Val		External validation of existing (previously developed) model in other participants

This table should be completed once for each publication being assessed and for each relevant outcome in your review.	
Publication reference	
Models of interest	
Outcome of interest	

Step 3: Assess risk of bias and applicability

PROBAST is structured as four key domains. Each domain is judged for risk of bias (low, high or unclear) and includes signalling questions to help make judgements. Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). All signalling questions are phrased so that "yes" indicates absence of bias. Any signalling question rated as "no" or "probably no" flags the potential for bias; you will need to use your judgement to determine whether the domain should be rated as "high", "low" or "unclear" risk of bias. The guidance document contains further instructions and examples on rating signalling questions and risk of bias for each domain.
The first three domains are also rated for concerns for applicability (low/ high/ unclear) to your review question defined above.
Complete all domains separately for each evaluation of a distinct model. Shaded boxes indicate where signalling questions do not apply and should not be answered.

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection:			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?			
1.2 Were all inclusions and exclusions of participants appropriate?			
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*		
Rationale of bias rating:			
B. Applicability			
Describe included participants, setting and dates:			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*		
Rationale of applicability rating:			

554

DOMAIN 2: Predictors			
A. Risk of Bias			
List and describe predictors included in the final model, e.g. definition and timing of assessment:			
		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?			
2.2 Were predictor assessments made without knowledge of outcome data?			
2.3 Are all predictors available at the time the model is intended to be used?			
Risk of bias introduced by predictors or their assessment	**RISK:** *(low/ high/ unclear)*		
Rationale of bias rating:			
B. Applicability			
Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** *(low/ high/ unclear)*		
Rationale of applicability rating:			

555

DOMAIN 3: Outcome			
A. Risk of Bias			
Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:			
		Dev	Val
3.1 Was the outcome determined appropriately?			
3.2 Was a pre-specified or standard outcome definition used?			
3.3 Were predictors excluded from the outcome definition?			
3.4 Was the outcome defined and determined in a similar way for all participants?			
3.5 Was the outcome determined without knowledge of predictor information?			
3.6 Was the time interval between predictor assessment and outcome determination appropriate?			
Risk of bias introduced by the outcome or its determination	**RISK:** *(low/ high/ unclear)*		
Rationale of bias rating:			
B. Applicability			
At what time point was the outcome determined:			
If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:			
Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** *(low/ high/ unclear)*		
Rationale of applicability rating:			

556

DOMAIN 4: Analysis
Risk of Bias

Describe numbers of participants, number of candidate predictors (for DEV only), outcome events and events per candidate predictor (for DEV only):

Describe how the model was developed (predictor selection, optimism, risk groups, model performance):

Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants):

Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit:

Describe any participants who were excluded from the analysis:

Describe missing data on predictors and outcomes as well as methods used for missing data:

		Dev	Val
4.1 Were there a reasonable number of participants with the outcome?			
4.2 Were continuous and categorical predictors handled appropriately?			
4.3 Were all enrolled participants included in the analysis?			
4.4 Were participants with missing data handled appropriately?			
4.5 Was selection of predictors based on univariable analysis avoided?			
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?			
4.7 Were relevant model performance measures evaluated appropriately?			
4.8 Was model overfitting and optimism in model performance accounted for?			
4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?			
Risk of bias introduced by the analysis	**RISK:** *(low/ high/ unclear)*		

Rationale of bias rating:

Step 4: Overall assessment

Use the following tables to reach overall judgements about risk of bias and concerns for applicability of the prediction model evaluation (development and/or validation) across all assessed domains.
Complete for each evaluation of a distinct model.

Reaching an overall judgement about risk of bias of the prediction model evaluation	
Low risk of bias	If all domains were rated low risk of bias. If a prediction model was developed without any external validation, and it was rated as low risk of bias for all domains, consider downgrading to **high risk of bias**. Such a model can only be considered as low risk of bias, if the development was based on a very large data set and included some form of internal validation.
High risk of bias	If at least one domain is judged to be at **high risk of bias**.
Unclear risk of bias	If an unclear risk of bias was noted in at least one domain and it was low risk for all other domains.

Reaching an overall judgement about applicability of the prediction model evaluation	
Low concerns for applicability	If low concerns for applicability for all domains, the prediction model evaluation is judged to have **low concerns for applicability**.
High concerns for applicability	If high concerns for applicability for at least one domain, the prediction model evaluation is judged to have **high concerns for applicability**.
Unclear concerns for applicability	If unclear concerns (but no "high concern") for applicability for at least one domain, the prediction model evaluation is judged to have **unclear concerns for applicability** overall.

Overall judgement about risk of bias and applicability of the prediction model evaluation		
Overall judgement of risk of bias	**RISK:** *(low/ high/ unclear)*	

Summary of sources of potential bias:

Overall judgement of applicability	**CONCERN:** *(low/ high/ unclear)*	

Summary of applicability concerns:

Members of PROBAST Delphi group

Members of PROBAST steering group

Robert F. Wolff, Kleijnen Systematic Reviews, York, United Kingdom
Karel G. M. Moons, Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Richard D. Riley, Keele University, United Kingdom
Penny F. Whiting, University Hospitals Bristol NHS Foundation Trust, United Kingdom; University of Bristol, United Kingdom
Marie Westwood, Kleijnen Systematic Reviews, York, United Kingdom
Gary S. Collins, Centre for Statistics in Medicine, NDORMS, University of Oxford, United Kingdom
Johannes B. Reitsma, Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Jos Kleijnen, Kleijnen Systematic Reviews, York, United Kingdom; School for Public Health and Primary Care (CAPHRI), Maastricht University, Maastricht, The Netherlands
Sue Mallett, Institute of Applied Health Sciences, University of Birmingham, United Kingdom

Members of PROBAST Delphi group (in alphabetical order)

Prof Doug Altman, PhD. Centre for Statistics in Medicine, NDORMS, University of Oxford, United Kingdom
Prof Patrick Bossuyt, PhD. Division Clinical Methods & Public Health, University of Amsterdam, The Netherlands
Prof Nancy R. Cook, ScD. Brigham and Women's Hospital, Boston, United States of America
Gennaro D´Amico, MD. Ospedale V Cervello, Palermo, Italy
Thomas P. A. Debray, PhD, MSc. Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Prof Jon Deeks, PhD. Institute of Applied Health Research, University of Birmingham, United Kingdom
Joris de Groot, PhD. Philips Image Guided Therapy Systems, Best, The Netherlands
Emanuele di Angelantonio, PhD, MSc. Department of Public Health and Primary Care, University of Cambridge, United Kingdom
Prof Tom Fahey, MD, MSc. Royal College of Surgeons in Ireland, Dublin, Ireland
Prof Frank Harrell, PhD. Department of Biostatistics, Vanderbilt University, United States of America
Prof Jill A. Hayden, PhD. Department of Community Health and Epidemiology, Dalhousie University, Canada
Martijn W. Heymans, PhD. Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, The Netherlands
Lotty Hooft, PhD. Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Prof Chris Hyde, PhD. Institute of Health Research, University of Exeter Medical School, United Kingdom
Prof John Ioannidis, MD, DSc. Meta-Research Innovation Center at Stanford (METRICS), Stanford University, United States of America
Prof Alfonso Iorio, MD, PhD. Department of Health Research Methods, Evidence, and Impact (HEI), McMaster University, Canada
Stephen Kaptoge, PhD. Department of Public Health & Primary Care, University of Cambridge, United Kingdom
Prof André Knottnerus, MD, PhD. Department of Family Medicine, Maastricht University, The Netherlands
Mariska Leeflang, PHD, DVM. Department of Clinical Epidemiology, Biostatistics and Boinformatics, University of Amsterdam, The Netherlands
Frances Nixon, BSc. National Institute for Health and Care Excellence (NICE), Manchester, United Kingdom
Prof Pablo Perel, MD, PhD, MSc. Centre for Global Chronic Conditions, London School of Hygiene and Tropical Medicine, United Kingdom
Bob Phillips, PhD, MMedSci. Centre for Reviews and Dissemination (CRD), York, United Kingdom
Heike Raatz, MD, MSc. Kleijnen Systematic Reviews, York, United Kingdom
Rob Riemsma, PhD. Kleijnen Systematic Reviews, York, United Kingdom
Prof Maroeska Rovers, PhD. Departments of Operating Rooms and Health Evidence, Radboud University Medical Center, Nijmegen, The Netherlands
Anne W. S. Rutjes, PhD, MHSc. Institute for Social and Preventive Medicine (ISPM) and Institute of Primary Health Care (BIHAM), University of Bern, Switzerland
Prof Willi Sauerbrei, PhD. Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany
Stefan Sauerland, MD, MPH. Institute for Quality and Efficiency in Healthcare (IQWiG), Cologne, Germany
Fülöp Scheibler, PhD, MA. University Medical Center Schleswig-Holstein, Kiel, Germany
Prof Rob Scholten, MD, PhD. Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Ewoud Schuit, PhD, MSc. Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, The Netherlands
Prof Ewout Steyerberg, PhD. Department of Public Health, Erasmus University Medical Center Rotterdam and Department of Biomedical Data Sciences, Leiden University Medical Center, The Netherlands
Toni Tan, MSc. National Institute for Health and Care Excellence (NICE), Manchester, United Kingdom
Gerben ter Riet, MD, PhD. Department of General Practice, University of Amsterdam, The Netherlands
Prof Danielle van der Windt, PhD. Centre for Prognosis Research, Keele University, United Kingdom
Yvonne Vergouwe, PhD. Department of Public Health, Erasmus University Medical Center, Rotterdam, The Netherlands
Andrew Vickers, PhD. Memorial Sloan-Kettering Cancer Center, New York, United States of America
Angela M. Wood, PhD. Department of Public Health and Primary Care, University of Cambridge, United Kingdom

PROBAST: A tool to assess the risk of bias and applicability of prediction model studies

Boxes

Box 1. Types of diagnostic and prognostic modelling studies or reports addressed by PROBAST

(adopted from the TRIPOD and CHARMS guidance(7, 26))

Prediction model development without external validation

These studies aim to develop one or more prognostic or diagnostic prediction models from a specific development data set. They aim to identify the important predictors of the outcome under study, assign weights (e.g. regression coefficients) to each predictor using some form of multivariable analysis, develop a prediction model to be used for individualised predictions, and quantify the predictive performance of that model in the development set. Sometimes, model development studies may also focus on adding one or more new predictors to established predictors. In any prediction model study, overfitting may occur, particularly in small data sets. Hence, development studies should include some form of resampling or "internal validation" (internal because the same data are used for both development and internal validation), such as bootstrapping or cross-validation. These methods quantify any optimism (bias) in the predictive performance of the developed model.

Prediction model development with external validation

Studies that have the same aim as the previous type, but the development of the model is followed by quantifying the model predictive performance in data *external* to the development sample i.e. from different participants. This may be data collected by the same investigators, commonly using the same predictor and outcome definitions and measurements, but sampled from a later time period (temporal validation); by other investigators in another hospital or country, sometimes using different definitions and measurements (geographic validation); in similar participants, but from an intentionally chosen different setting (e.g. model developed in secondary care and tested in similar participants from primary care); or even in other types of participants (e.g. model developed in adults and tested in children). Randomly splitting a single data set into a development and a validation data set is often erroneously referred to as a form of external validation, but actually is an inefficient form of "internal" validation, because the two so created data sets only differ by chance and sample size of model development is reduced**.**

When a model predicts poorly when validated in other data, a model validation can be followed by adjusting (or updating the existing model (e.g. by recalibration of the baseline risk or hazard or adjusting the weights of the predictors in the model) to the validation data set at hand, and even by extending the model by adding new predictors to the existing model. In both situations in fact a new model is being developed after the external validation of the existing model.

Prediction model external validation

These studies aim to assess the predictive performance of one or more existing prediction models by using in data *external* to the development sample i.e. from different participants.

Box 2. Differences between diagnostic and prognostic prediction model studies

Diagnostic prediction models aim to estimate the probability that a target condition measured using a reference standard (referred to as outcome in PROBAST) is currently present or absent within an individual. In diagnostic prediction model studies, the prediction is for an outcome already present so the preferred design is a cross-sectional study although sometimes follow-up is used as part of the reference test to determine the target condition presence at the moment of prediction.

Prognostic prediction models estimate whether an individual will experience a specific event or outcome in the future within a certain time period, ranging from minutes to hours, days, weeks, months or years: always a longitudinal relationship.

Despite the different timing of the predicted outcome, there are many similarities between diagnostic and prognostic prediction models, including the:

- Type of outcome is often binary (target condition or disease presence (yes/no) or future occurrence of an outcome event (yes/no).
- Key interest is to estimate the probability of an outcome being present or occurring in the future based on multiple predictors with the purpose of informing individuals and guiding decision-making.
- Same challenges occur when developing or validating multivariable prediction models. The same measures for assessing predictive performance of the model can be used, although diagnostic models more frequently extend assessment of predictive performance to focus on thresholds of clinical relevance.

There are also various differences in terminology between diagnostic and prognostic model studies:

Diagnostic prediction model study	**Prognostic prediction model study**
Predictors	
Diagnostic tests or index tests	Prognostic factors or prognostic indicators
Outcome	
Reference standard used to assess or verify presence/absence of target condition	Event (future occurrence yes or no) Event measurement
Missing outcome assessment	
Partial verification, lost to follow-up	Lost to follow-up and censoring

Tables

Table 1. Four steps in PROBAST

Step	Task	When to complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability (per domain)	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment of risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication

.

Table 2. Summary of step 3 (Assessment of risk of bias and concerns for applicability)

	1. Participants	2. Predictors	3. Outcome	4. Analysis
Signalling questions	1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	2.1 Were predictors defined and assessed in a similar way for all participants?	3.1 Was the outcome determined appropriately?	4.1 Were there a reasonable number of participants with the outcome?
	1.2 Were all inclusions and exclusions of participants appropriate?	2.2 Were predictor assessments made without knowledge of outcome data?	3.2 Was a pre-specified or standard outcome definition used?	4.2 Were continuous and categorical predictors handled appropriately?
		2.3 Are all predictors available at the time the model is intended to be used?	3.3 Were predictors excluded from the outcome definition?	4.3 Were all enrolled participants included in the analysis?
	–		3.4 Was the outcome defined and determined in a similar way for all participants?	4.4 Were participants with missing data handled appropriately?
	–	–	3.5 Was the outcome determined without knowledge of predictor information?	4.5 Was selection of predictors based on univariable analysis avoided? [D]
	–	–	3.6 Was the time interval between predictor assessment and outcome determination appropriate?	4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?
	–	–		4.7 Were relevant model performance measures evaluated appropriately?
	–	–	–	4.8 Was model overfitting, underfitting and optimism in model performance accounted for? [D]
	–	–	–	4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis? [D]
ROB	Selection of participants	Predictors or their assessment	Outcome or its determination	Analysis
Applicability	Included participants and setting do not match the review question	Definition, assessment or timing of predictors in the model do not match the review question	Outcome, its definition, timing or determination do not match the review question	–

For further details please refer to [REF M18-1377] and www.probast.org
Signalling questions are rated as yes (Y), probably yes (PY), probably no (PN), no (N) or no information (NI). Risk of bias and concerns for applicability are rated as low, high, or unclear.
D = Development studies only; ROB = Risk of bias; V = Validation studies only

Table 3. Suggested Tabular Presentation for PROBAST Results

Study	Risk of bias				Applicability			Overall	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	Risk of bias	Applicability
Study 1	+	-	?	+	+	+	+	-	+
Study 2	+	+	+	+	+	+	+	+	+
Study 3	+	+	+	?	-	+	+	?	-
Study 4	-	?	?	-	+	+	-	-	-
Study 5	+	+	+	+	+	?	+	+	?
Study 6	+	+	+	+	?	+	?	+	?
Study 7	?	?	+	?	+	+	+	?	+
Study 8	+	+	+	+	+	+	+	+	+

+ = low risk of bias / low concerns regarding applicability; - = high risk of bias / high concerns regarding applicability; ? = unclear risk of bias / unclear concerns regarding applicability